Drink. Water Eng. Sci., 10, 75–82, 2017
https://doi.org/10.5194/dwes-10-75-2017

Drinking Water
Engineering and Science
Open Access

Modeling and clustering water demand patterns from real-world smart meter data

Nicolas Cheifetz[1], Zineb Noumir[2], Allou Samé[2], Anne-Claire Sandraz[1], Cédric Féliers[1], and Véronique Heim[3]

[1]Veolia Eau d'Ile de France, Le Vermont, 28, Boulevard de Pesaro, Nanterre 92751, France
[2]Université Paris-Est, IFSTTAR, COSYS, GRETTIA, Marne-la-Vallée 77447, France
[3]Syndicat des Eaux d'Ile de France (SEDIF), 120 Boulevard Saint-Germain, Paris 75006, France

Correspondence to: Nicolas Cheifetz (nicolas.cheifetz@veolia.com)

Received: 19 March 2017 – Discussion started: 23 March 2017
Revised: 23 June 2017 – Accepted: 3 July 2017 – Published: 18 August 2017

Abstract. Nowadays, drinking water utilities need an acute comprehension of the water demand on their distribution network, in order to efficiently operate the optimization of resources, manage billing and propose new customer services. With the emergence of smart grids, based on automated meter reading (AMR), a better understanding of the consumption modes is now accessible for smart cities with more granularities. In this context, this paper evaluates a novel methodology for identifying relevant usage profiles from the water consumption data produced by smart meters. The methodology is fully data-driven using the consumption time series which are seen as functions or curves observed with an hourly time step. First, a Fourier-based additive time series decomposition model is introduced to extract seasonal patterns from time series. These patterns are intended to represent the customer habits in terms of water consumption. Two functional clustering approaches are then used to classify the extracted seasonal patterns: the functional version of K-means, and the Fourier REgression Mixture (FReMix) model. The K-means approach produces a hard segmentation and K representative prototypes. On the other hand, the FReMix is a generative model and also produces K profiles as well as a soft segmentation based on the posterior probabilities. The proposed approach is applied to a smart grid deployed on the largest water distribution network (WDN) in France. The two clustering strategies are evaluated and compared. Finally, a realistic interpretation of the consumption habits is given for each cluster. The extensive experiments and the qualitative interpretation of the resulting clusters allow one to highlight the effectiveness of the proposed methodology.

1 Introduction

All modern cities need to deal with increasing populations and climate change while maintaining adequate water services for consumers. Here, climate change is only mentioned to emphasize the systemic changes inherent in any smart city. Until now, water or energy consumption readings have traditionally been collected once or twice a year in large territories (for example, regions or nations). With the arrival of smart grid meters, this situation has changed, and indexes can now be collected automatically with more granularities. The management of smart cities (Giffinger et al., 2007; Nam and Pardo, 2011) is based on automated electronic meters that are deployed on the distribution network and are used to handle billing and customer services. The first researches performed in the area of demand patterns classification belong to the electricity network fields (Irwin et al., 1986; Hernández et al., 2012). Most of the research in the water field is focused on demand forecasting (Donkor et al., 2012). Several approaches have been proposed for this purpose, including statistical forcasting models (Adamowski, 2008; Blokker et al., 2009). The emergence of smart meters shifts this research to classification of water demand (Aksela and Aksela, 2010). McKenna et al. (2014) proposed a procedure for classification of water demands recorded from smart meters us-

ing a Gaussian mixture model for feature selection and then the classical K-means algorithm for clustering (MacQueen, 1967).

In various applications, the data to be analyzed are not multivariate observations, but these can be seen as functions or curves that are either continuous or discrete, namely functional data. Such studies usually refer to functional data analysis (FDA) when data are varying in a continuum and potentially infinite dimensional (Ramsay and Silverman, 2005; Wang et al., 2015). Examples of functional data encompass longitudinal data, responses in medical treatments and objects in video sequences.

In the current case of smart meters, each signal is seen as a temporal function and is collected intermittently at discrete time points. Analyzing smart meter consumption is useful for water utilities in order to develop innovative capabilities in terms of grid management, planning and customer services. Functional clustering aggregates data mining techniques, which aim to identify homogeneous groups among functional data without using prior knowledge about their group labels (unknown cluster membership). Aiming to analyze household consumption, Cardell-Oliver (2013) introduces a methodology to cluster daily water use signature patterns based on expert rules and a classical K-means. Many functional clustering methods have been developed over the last decade. These methods can usually be separated into two categories: nonparametric methods using specific distances or dissimilarities between curves (Dabo-Niang et al., 2007), and mixture-model-based methods (Samé et al., 2011; Jacques and Preda, 2014). The collected curves can be multivariate, leading to a large representation space like in (Cheifetz et al., 2013) for change-point detection based on a specific curve modeling. The approach of the regression mixture model proposed by Gaffney and Smyth (2003) motivated the focus of this article.

This paper is organized as follows: the overall methodology is described in Sect. 2. This methodology is decomposed into two consecutive steps, that is to say, the extraction of seasonal patterns from time series in Sect. 3, and the identification of clusters with their profiles in Sect. 4 based on two clustering strategies: a functional version of K-means and a dedicated expectation maximization (EM) algorithm. Sect. 5 introduces the experimental data set, and an analysis of the clustering results is given. Finally, the article ends with a conclusion and some perspectives.

2 General methodology

The aim of this paper is to identify automatically the major water usage patterns in a set of time series recorded by smart water meters. A multi-step methodology is formulated to address this problem, as illustrated in Fig. 1. The first step consists in extracting the seasonal part of each time series, which represents the habits of water consumption for each meter, using a Fourier-based time series decomposition. Then, these seasonal components are normalized and used as input data by clustering algorithms. Two algorithms are used to classify the functional data into various water usage clusters. The first one consists in using the K-means jointly with the functional principal component analysis (FPCA) method, and the second one is based on a Fourier regression mixture model recently introduced by Samé et al. (2016). Both the seasonal extraction and clustering approaches are described in the next sections.

3 Extracting seasonal patterns from time series

Let $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ denote n time series, where each one of them, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})$, corresponds to hourly consumptions recorded by a single water meter, that is to say, $\boldsymbol{y}_i$ is a univariate time series and y_{it} is a real-valued scalar. It is implicitly assumed that all the series are recorded over the same time grid indexed by the ordered times $\{1, \ldots, T\}$ for all n curves.

3.1 Fourier-based time series decomposition

The methodology developed in this paper is based on the following classical additive decomposition:

$$y_{it} = f_{it} + x_{it} + d_{it} + \varepsilon_{it}, \tag{1}$$

where

- f_{it} is the global trend of the time series which is modeled in a nonparametric way using moving averages (Gourieroux et al., 1997), and
- x_{it} is the seasonal component. As the studied water consumption time series are subject to daily and weekly seasonality, a Fourier basis decomposition (De Livera et al., 2011) is formulated:

$$x_{it} = \sum_{j=1}^{q_1} \left[\alpha_j^{(1)} \cos\left(\frac{2\pi jt}{24}\right) + \alpha_j^{(2)} \sin\left(\frac{2\pi jt}{24}\right) \right] + \sum_{j=1}^{q_2} \left[\alpha_j^{(3)} \cos\left(\frac{2\pi jt}{168}\right) + \alpha_j^{(4)} \sin\left(\frac{2\pi jt}{168}\right) \right], \tag{2}$$

  where q_1 and q_2 are the respective numbers of trigonometric terms used to handle the daily and weekly seasonality, and $\alpha_j^{(1)}, \alpha_j^{(2)}, \alpha_j^{(3)}$, and $\alpha_j^{(4)}$ are the coefficients to be estimated. This trigonometric modeling has the advantage of requiring considerably fewer parameters compared to an approach based on dummy variables (De Livera et al., 2011).
- d_{it} is a component devoted to capturing the effect of exceptional public non-working days in France (e.g.,

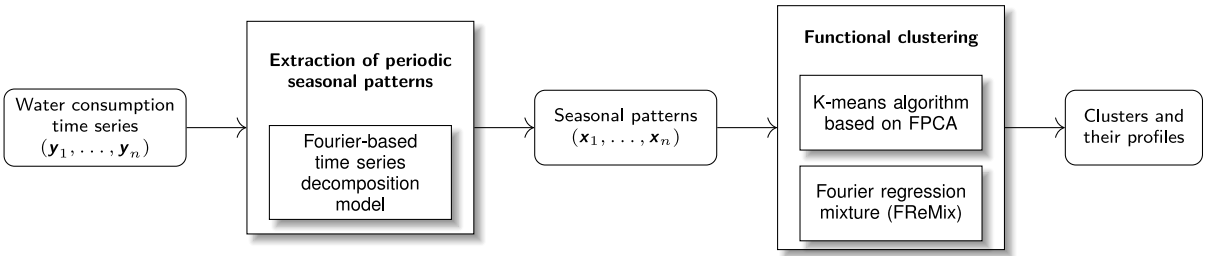

Figure 1. Block diagram describing the global methodology.

1 January, 1 May, Christmas Day). The following decomposition is used: $d_{it} = \sum_{j=1}^{24} \gamma_j \delta_{tj}$, where $\delta_{tj} = 1$ if t corresponds to the hour j of a non-working day and $\delta_{tj} = 0$ otherwise.

– ε_{it} is a centered Gaussian noise.

For compliance with the additivity and Gaussianity assumptions of this decomposition model, each time series $(y_{it})_{t=1,\ldots,T}$ was replaced by $(\log(y_{it} + \lambda))_{t=1,\ldots,T}$, where λ is a small positive number preventing degeneracy caused by null consumptions. Note that this transformation is used in the same way as the well-known Box–Cox transformation (Box and Cox, 1964).

3.2 Parameter estimation and practical use of the model

Given a time series $\boldsymbol{y}_i$ recorded by a smart meter, the trend $\boldsymbol{f}_i$ is estimated using a simple moving average (Gourieroux et al., 1997; Shumway and Stoffer, 2010). As the daily and weekly periodicities (24 and 168) should be removed from the univariate time series, a centered moving average of order 168 is performed.

After estimating the trend and given a couple (q_1, q_2), the coefficients α_{1j}, α_{2j}, α_{3j}, α_{4j} and γ_j are simultaneously identified by performing a multiple linear regression of $(y_{it} - f_{it})$ over the variables $\cos\left(\frac{2\pi jt}{24}\right)$, $\sin\left(\frac{2\pi jt}{24}\right)$, $\cos\left(\frac{2\pi jt}{168}\right)$, $\sin\left(\frac{2\pi jt}{168}\right)$ and δ_{tj}. Selecting the couple (q_1, q_2) remains a sensitive point which can ideally be addressed by choosing the couple which optimizes a model selection criterion such as the Akaike information criterion (AIC) introduced by Akaike (1974) or the Bayesian information criterion (BIC) introduced by Schwarz (1978). In this paper, several combinations of (q_1, q_2) were tested and the couple $(4, 24)$ has been selected, leading to a good compromise between visual representation of seasonal patterns and modeling accuracy. An example of decomposition of a time series is shown in Fig. 2. The trend is displayed together with the complete time series, while the seasonal component is displayed with the weekly sub-series.

From each time series $\boldsymbol{y}_i$, the model parameters defined by Eq. (1) are thus identified, and the periodic seasonal pattern defined by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$, with $m = 168$, is extracted. Due to the periodicity of the series $(x_{it}, \ldots, x_{iT})$ defined by Eq. (2), it should be noted that the first terms $m = 168$ are sufficient to characterize the time series. Then, the set of seasonal patterns $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is standardized as suggested by Gaffney (2004): $x_{it} \leftarrow \dfrac{x_{it} - (1/m)\sum_{j=1}^{m} x_{ij}}{\sigma(\boldsymbol{x}_i)}, \forall i, t$, where $\sigma(\boldsymbol{x}_i)$ is the standard deviation of $\boldsymbol{x}_i$. The set of normalized seasonal patterns is used as input data for the clustering step which will be described in the following section.

It is worth noting that the proposed decomposition can also be used to fill missing values that may occur along the time series. The reconstruction formula is $\hat{y}_{it} = \hat{f}_{it} + \hat{x}_{it} + \hat{d}_{it}$, where $\hat{f}_{it}$, $\hat{x}_{it}$, and $\hat{d}_{it}$ are the estimated components.

4 Clustering seasonal profiles

In order to extract relevant usage profiles from water consumption time series, two functional data clustering approaches are considered in this paper: the first one is the functional version of the K-means algorithm and the second one is based on a specific Fourier regression mixture model.

4.1 Functional clustering based on FPCA

In this subsection, the clustering method (Peng and Müller, 2008; Sood et al., 2009) is inspired by functional data analysis (Ramsay and Silverman, 2005; Wang et al., 2015) which assumes that data are functions or curves. This clustering approach is mainly based on functional principal component analysis (FPCA) and can be summarized in the following two consecutive steps.

Smoothing and dimension reduction This step consists in converting the n time series $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ into functional objects $(\boldsymbol{x}_1(t), \ldots, \boldsymbol{x}_n(t))$ and then applying the classical PCA to the multivariate data obtained by discretizing the functions $\boldsymbol{x}_i(t)$ over the temporal grid $\{1, \ldots, m\}$. In this paper, the PCA is directly performed on the seasonal patterns $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ that are based on trigonometric (smooth) functions, and the principal components are selected such that 95 % of the data variance is explained.

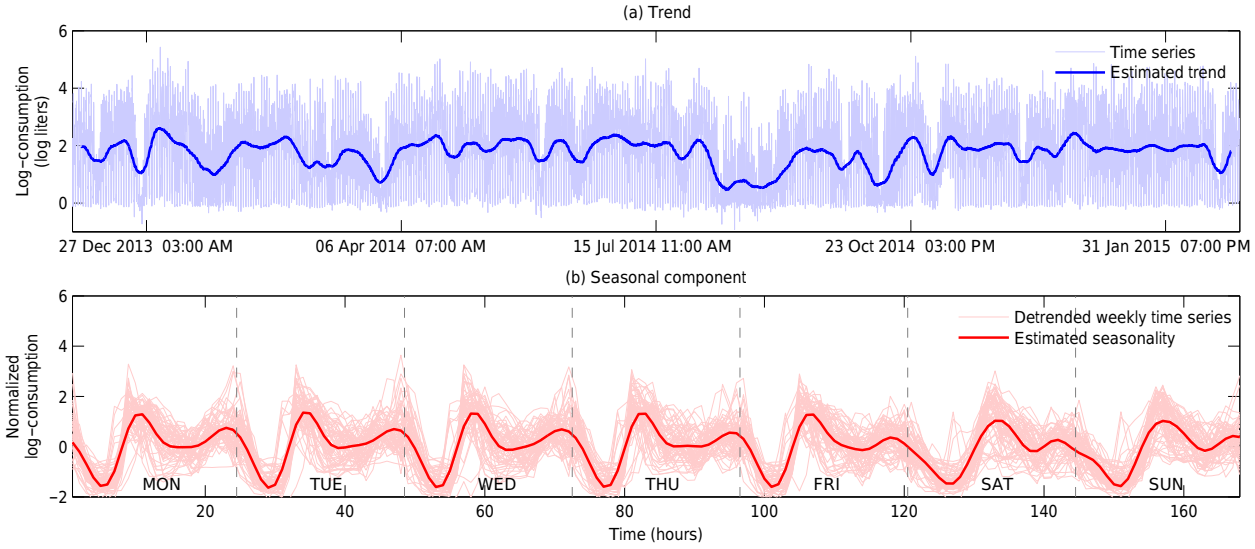

Figure 2. Extraction of periodic seasonal patterns using Fourier-based time series decomposition. The trend is displayed with the complete time series **(a)** and the seasonal component is displayed with the weekly time series **(b)**.

Clustering in this step consists of a classical clustering method performed on the principal component scores estimated previously. The well-known K-means algorithm (MacQueen, 1967) is applied using several random initializations and the partition with the lowest intra-cluster inertia is selected.

The resulting functional clustering strategy is called FPCA-KM. The number of clusters K has been selected by minimizing the BIC-like penalized criterion $\mathrm{BIC}(K) = C + \nu_K \log(n)$, where C is the intra-cluster inertia optimized by the K-means algorithm and $\nu_K = Kq$ is the number of parameters to be estimated with q the number of selected principal components.

The general idea of PCA is to create a small number of uncorrelated variables with maximal variance. The extension of this technique for functional data is proposed in the work of Ramsay and Silverman (2005) and Ferraty and Vieu (2006). The FPCA is an efficient tool providing common functional components explaining the structures of individual trajectories.

4.2 Fourier regression mixture model

Inspired by the polynomial regression mixture model formulated by Gaffney and Smyth (1999), this subsection introduces a Fourier regression mixture model, called the FReMix model. The Fourier regression mixture was preferred to polynomial and spline regression mixtures for its compliance with the modeling adopted in the first step (seasonal pattern extraction). Moreover, a Fourier polynomial is a universal approximator of functions and remains a good candidate in modeling clusters whose prototypes are nonlinear and potentially periodic functions.

4.2.1 Model definition

Unlike standard vector-based mixture models, the density of each component of the FReMix model is represented by a trigonometric prototype function that is parameterized by regression coefficients and a noise variance. The prototype functions represent the class conditional expectations of $\boldsymbol{x}_i$. The Fourier regression mixture model therefore assumes that each time series $\boldsymbol{x}_i$ is distributed according to the following density:

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_i;\mathbf{U}\boldsymbol{\beta}_k,\sigma_k^2\mathbf{I}\right), \tag{3}$$

where $\boldsymbol{\theta} = (\pi_1,\ldots,\pi_K,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K,\sigma_1^2,\ldots,\sigma_K^2)$ is the complete parameter vector. The probabilities π_k are the proportions of the mixture satisfying $\sum_{k=1}^{K}\pi_k = 1$, $\boldsymbol{\beta}_k = (\beta_{k,1},\ldots,\beta_{k,2(q_1+q_2)})' \in \mathbb{R}^{2(q_1+q_2)}$ is the coefficient vector of the kth regression model and $\sigma_k^2 > 0$ is the associated noise variance. The matrix $\mathbf{U} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_m)'$ is a regression matrix of size $m \times 2(q_1+q_2)$, where the vector $\boldsymbol{u}_t \in \mathbb{R}^{2(q_1+q_2)}$ is defined by $(\forall t = 1,\ldots,m)$:

$$\boldsymbol{u}_t = \left[\cos\left(\frac{2\pi t}{24}\right)\ \sin\left(\frac{2\pi t}{24}\right)\ \cdots\ \cos\left(\frac{2\pi q_1 t}{24}\right)\ \sin\left(\frac{2\pi q_1 t}{24}\right)\right.$$
$$\left.\cos\left(\frac{2\pi t}{168}\right)\ \sin\left(\frac{2\pi t}{168}\right)\ \cdots\ \cos\left(\frac{2\pi q_2 t}{168}\right)\ \sin\left(\frac{2\pi q_2 t}{168}\right)\right]',$$

and $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. This specific mixture model corresponds to the class-specific prototype functions $g_k(t) = \boldsymbol{\beta}_k'\boldsymbol{u}_t$

which is also given by

$$g_k(t) = \sum_{j=1}^{q_1}\left[\beta_{k,2j-1}\cos\left(\frac{2\pi jt}{24}\right)+\beta_{k,2j}\sin\left(\frac{2\pi jt}{24}\right)\right] + \sum_{j=1}^{q_2}\left[\beta_{k,2q_1+2j-1}\cos\left(\frac{2\pi jt}{168}\right) + \beta_{k,2q_1+2j}\sin\left(\frac{2\pi jt}{168}\right)\right]. \quad (4)$$

4.2.2 EM algorithm and practical issues

Assuming that the n seasonal time series $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ are independent, the parameter vector $\boldsymbol{\theta}$ is estimated in the same way as for the classical Gaussian mixture model (McLachlan and Krishnan, 2007) and the polynomial regression mixture model (Gaffney and Smyth, 1999), by maximizing the specific log-likelihood $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{x}_i;\mathbf{U}\boldsymbol{\beta}_k,\sigma_k^2\mathbf{I}\right)$ via the EM procedure (Dempster et al., 1977; Gaffney and Smyth, 1999; McLachlan and Krishnan, 2007). The pseudo-code can be found in the paper by Samé et al. (2016). As a reminder, the couple $(q_1,q_2)=(4,24)$ is selected in the seasonal pattern extraction step (cf. Sect. 3.2). The algorithm is initialized as follows: the initial regression coefficients and variances are obtained by performing a Fourier regression separately on K seasonal series randomly drawn into the data set $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ and the initial proportions of the latent classes are set to $\pi_k=\frac{1}{K}$. This process is repeated 20 times and the parameters with the highest log-likelihood are selected. The number of clusters is selected through the BIC criterion (Schwarz, 1978) defined by $\mathrm{BIC}(K) = -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + \nu_K\log(n)$, where $\hat{\boldsymbol{\theta}}$ is the parameter vector estimated by the EM algorithm, and ν_K is the number of free parameters of the model: $\nu_K = 2K(q_1+q_2+1)-1$.

After estimating the parameter vector $\boldsymbol{\theta}$, a time series partition is obtained by assigning each series $\boldsymbol{x}_i$ to the cluster having the highest posterior probability $\tau_{ik} = \frac{\pi_k\mathcal{N}(\boldsymbol{x}_i;\mathbf{U}\boldsymbol{\beta}_k,\sigma_k^2\mathbf{I})}{\sum_{\ell=1}^{K}\pi_\ell\mathcal{N}(\boldsymbol{x}_i;\mathbf{U}\boldsymbol{\beta}_\ell,\sigma_\ell^2\mathbf{I})}$.

5 Experimental study using real data

5.1 Description of the data set

The experimental data set represents the water consumption recorded by a few smart meters deployed on the network of the Syndicat des Eaux d'Ile-de-France (SEDIF). The SEDIF is a large association including 150 municipalities which provides drinking water for more than 4 million inhabitants of suburban Paris. This is the largest drinking water distribution network (WDN) in France, with about 8000 km of pipes and more than 750 000 m^3 of water produced each day. The consumption is measured hourly (in liters) by 10 233 m during 15 months (from November 2013 to March 2015). The resulting data set is then made of univariate time series $(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)$, where $n=10\,233$ and the length of each time series $\boldsymbol{y}_i$ is $T=11\,016$. After the extraction of periodic seasonal patterns (cf. Sect. 3), a new set of time series $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ is built where the length of each seasonal pattern $\boldsymbol{x}_i$ is $m=168$. These series are used as input data for the clustering algorithms.

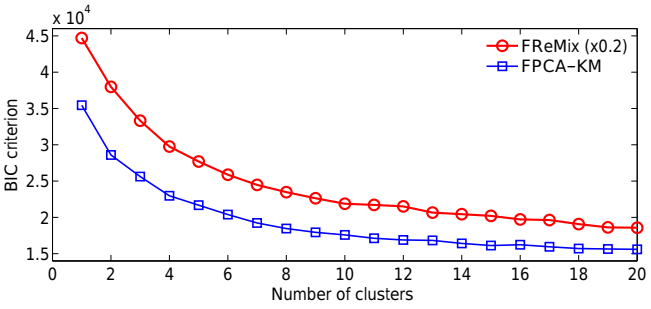

Figure 3. Evolution of the BIC criteria according to the number of clusters.

5.2 Selecting the number of clusters

The number of clusters for the two methods was selected by running the algorithms with several values of K and then choosing the value which minimizes the BIC criterion. Figure 3 shows the evolution of this criterion for the two clustering algorithms in relation to the number of clusters. For both methods, the BIC criterion exhibits a decrease continuously, while the K value increases. Nevertheless, it can be seen that the variation of BIC is not significant when the number of clusters is above eight. Therefore, the number of clusters is selected such that $K=8$.

5.3 Results interpretation and discussion

The seasonal time series are classified into $K=8$ clusters, using functional K-means (FPCA-KM strategy) and Fourier regression mixture (FReMix model) as illustrated, respectively, by Fig. 4a and b. For each cluster, the weekly prototype is displayed in orange (sub-figures on the left). Moreover, the right plots of Fig. 4a and b display the cluster profiles using a daily representation, the colors (from blue to red) indicating the day of the week (from Monday to Sunday). The percentage of input time series belonging to each cluster is also provided.

It can be observed that the consumption profiles are quite similar for the two methods, despite the differences in the cluster percentages. As no socio-demographic data about customers were available at this stage of the study, a qualitative evaluation of the results is performed and the pattern repartition shown in Fig. 4 can be explained by the following realistic categories.

Residential use Clusters 1, 2, and 3. The temporal dynamic of these clusters corresponds to customers who wake up

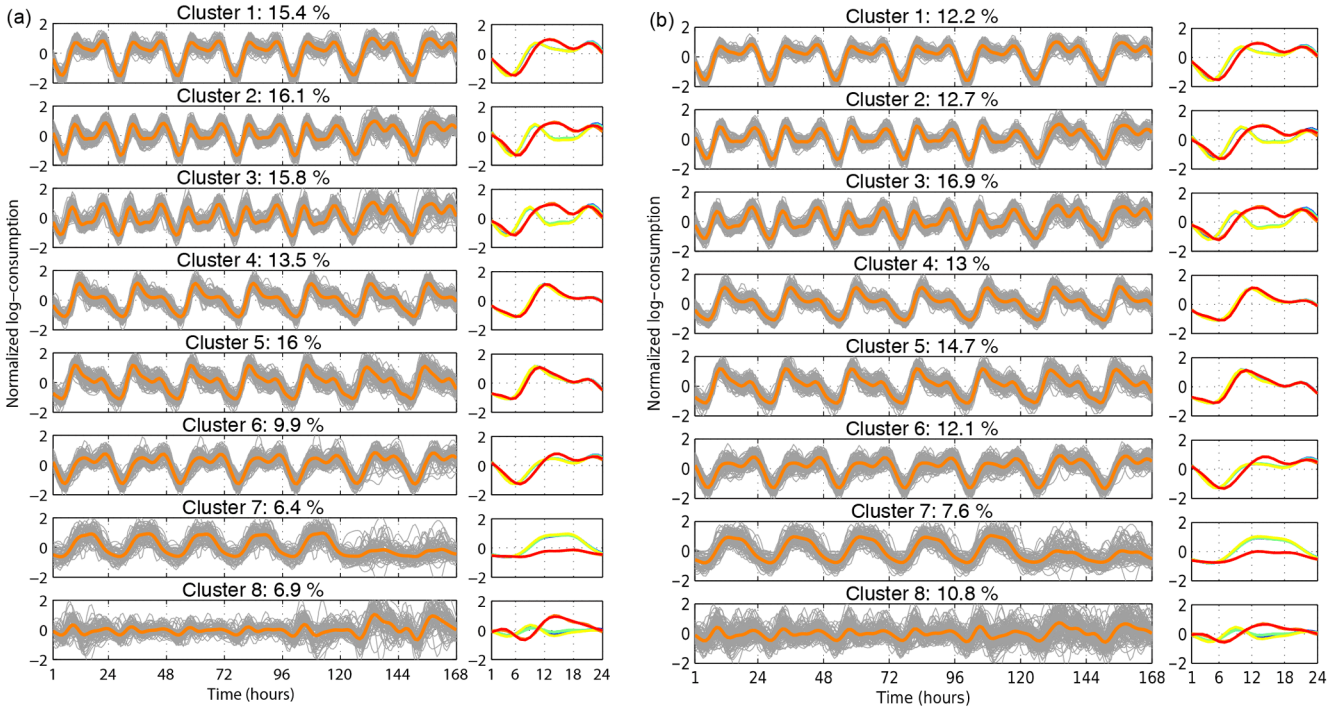

Figure 4. Clustering results obtained with the FPCA-KM **(a)** and the FReMix **(b)**. For each side, the subfigures on the left represent a weekly view of the clusters with their prototypes displayed in orange. The subfigures on the right are daily prototypes resulting from the segmentation of the weekly orange curves and colors (from blue to yellow to red) indicate the day of the week (from Monday to Sunday).

between 06:00 and 08:00, take a shower and then go to work. This habit is characterized by a consumption peak around 10:00 in the morning. The other peak, observed in the evening at around 20:00, corresponds to the return home. The minimum consumption level between these two peaks can be attributed to persons in households who stay at home during working hours.

Commercial use Clusters 4, 5, and 6. This category corresponds to a set of customers whose consumption habits are the same during working days and weekends. It may correspond, for example, to small businesses or medical centers that stay open every day and have the same daily consumption profile. It should be noticed that clusters 4 and 5 differ from the other clusters by their smaller evening peak.

Office or industrial use Cluster 7. One can observe an active water consumption from Monday to Friday (workdays) during the business hours, and a very low consumption during the weekend.

Noise cluster Cluster 8. This cluster, which has the largest variance, groups a set of atypical patterns which does not match with the other clusters. It can be considered a noise cluster.

Note that a functional clustering scheme is adopted because this is suitable for dealing with the analysis of our consumption curves. Indeed, these real-valued data can be seen as the realizations of a one-dimensional stochastic process, recorded on the same time grid (hourly spaced) of ordered times. In practice, data frames are frequently sent by modules which are physically connected to the meters; each consumption time series can be re-constructed based on a sequence of the data frames. Exogenous variables (e.g., weather inputs or meter localization) are not considered in this work due to a non-significant improvement in the results, but might be used in a future work.

Furthermore, the time series $\boldsymbol{x}_i$ have a length of 168 due to the trigonometric modeling of the chosen Fourier basis decomposition. The Fourier coefficients are identified by performing a multiple linear regression on the detrended global time series $\boldsymbol{y}_i$, which limits the effect of a long seasonality. Finally, our sample has a duration of 15 months (with five seasons at most), and we are more interested in identifying the major mode of consumption than estimating water demand profiles with local changes and a fine granularity.

However, we wanted to integrate some prior knowledge about day/week seasonality and exceptional public non-working days. A Fourier-based decomposition has the capacity to easily take into account this prior knowledge and this decomposition is consistent with our probabilistic FReMix model definition. A wavelet-based analysis could also be used for decomposition (keeping some local properties along temporal patterns), but integrating such prior knowledge

might not be straightforward and the number of parameters in this case should not be reduced significantly. In this article, we evaluate a probabilistic method and a geometrical approach. This second method is based on a K-means and minimizes the intra-cluster inertia which can be seen as an aggregated distance over the water time series. To our knowledge, the complexity of time series distances (e.g., dynamic time warping) can be prohibitive with a large time series data set.

This paper deals with an unsupervised classification problem based on water consumption time series. Water demand forecasting is not the issue in this paper; nevertheless, the resulting segmentation of water consumption time series can be used for several scientific problems, including sequential detection, predictive classification or demand forecasting. We assume no supervision in our setting due to a partial and uncertain knowledge of the usage labels; users do not inform systematically their water utilities when businesses change or people come in/leave a home. This is why there is no quantitative accuracy about clustering. Each log-consumption time series is standardized before clustering, which leads to a discrimination in terms of seasonal patterns and is not based on water volume. This explains why we called cluster 1 "office and industrial usage". Of course, industrial usage might produce erratic water patterns which would be classified in cluster 2. Partitioning the eight clusters into four categories would suggest non-negligible variations in residential use as well as commercial use, and extra investigations about the users which should not be underestimated in terms of time and cost. The EM algorithm used to fit the Fourier REgression Mixture (FReMix) model is flexible and can be reformulated in a future work with a semi-supervision (by fixing a set of posterior probabilities) or a partial supervision (e.g., using belief functions).

Identifying the major usage profiles from water consumption is an interesting topic to water utilities. Indeed, the resulting segmentation helps the water companies to gain better knowledge about users consuming the distributed water. The user has a better experience with the tools developed by their water utility. For instance, users at Veolia Eau d'Ile de France (Paris area in France) can already monitor their water index/consumption on a dedicated website for free. Using our clustering results, people could compare with similar patterns and adapt their consumptions according to their needs. In addition, the resulting clusters are used by an early warning system which alerts the user when a leakage occurs into the private network. An erratic water pattern (like in cluster 2) can be a sign of a leakage and might initiate a corrective action. Concerning the grid management, each prototype can be used to represent the water behavior of users belonging to the same cluster. Sampling a large amount of water meters is useful for several topics (e.g., tracking the meter metrology, estimating the global consumption modes based on a limited number of meters); such sampling analysis is straightforward using our meter segmentation.

6 Conclusions and perspectives

A general methodology is introduced in this paper for automatically discriminating several water usages and extracting relevant water consumption profiles from time series recorded by smart meters. Considering that the consumption habits are of interest and not the consumption levels, the first step of the method consists in extracting the seasonal part of time series using an additive classical decomposition model. This modeling of the seasonal component is based on a specific Fourier expansion which takes into account daily and weekly periodicities. As this study aims to identify relevant water usage profiles, two functional clustering techniques are used to classify the seasonal patterns extracted from the water consumption time series: a functional variant of the K-means algorithm and a specific EM algorithm based on a Fourier regression mixture model (FReMix). The FReMix model is richer than the other clustering approach in that the Fourier basis decomposition is fully integrated in the modeling and each cluster is described by its first two moments, while the K-means only extracts the mean curves. Furthermore, the K-means produces a hard segmentation, while the FReMix creates a soft partition where each cluster membership is weighted by a posterior probability. Eight clusters are then identified for the two clustering methods. The resulting prototypes are quite similar for the two approaches and a realistic category is given to each cluster.

More investigations are in progress with the water utility Veolia Eau d'Ile de France in order to refine the clustering results and the proposed methodology is also being applied to a new large-scale database.

Data availability. The data set is private and belongs to the SEDIF.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Computing and Control for the Water Industry, CCWI 2016". It is a result of the 14th International CCWI Conference, Amsterdam, the Netherlands, 7–9 November 2016.

Acknowledgements. The study presented in the paper was part of a previous work between IFSTTAR and Veolia Eau d'Ile de France. This work is now part of French–German collaborative research project ResiWater (http://www.resiwater.eu/) that is funded by the French National Research Agency (ANR; project: ANR-14-PICS-0003) and the German Federal Ministry of Education and Research (BMBF; project: BMBF-13N13690).

Edited by: Edo Abraham
Reviewed by: three anonymous referees

References

Adamowski, J. F.: Peak daily water demand forecast modeling using artificial neural networks, J. Water Res. Pl.-ASCE, 134, 119–128, 2008.

Akaike, H.: A new look at the statistical model identification, IEEE T. Automat. Contr., 19, 716–723, 1974.

Aksela, K. and Aksela, M.: Demand estimation with automated meter reading in a distribution network, J. Water Res. Pl.-ASCE, 137, 456–467, 2010.

Blokker, E., Vreeburg, J., and Van Dijk, J.: Simulating residential water demand with a stochastic end-use model, J. Water Res. Pl.-ASCE, 136, 19–26, 2009.

Box, G. E. and Cox, D. R.: An analysis of transformations, J. Roy. Stat. Soc. B, 211–252, 1964.

Cardell-Oliver, R.: Water use signature patterns for analyzing household consumption using medium resolution meter data, Water Resour. Res., 49, 8589–8599, 2013.

Cheifetz, N., Samé, A., Aknin, P., De Verdalle, E., and Chenu, D.: A Sequential Testing Procedure for Multiple Change-Point Detection in a Stream of Pneumatic Door Signatures, in: 12th International Conference on Machine Learning and Applications (ICMLA), IEEE, 117–122, 2013.

Dabo-Niang, S., Ferraty, F., and Vieu, P.: On the using of modal curves for radar waveforms classification, Comput. Stat. Data An., 51, 4878–4890, 2007.

De Livera, A. M., Hyndman, R. J., and Snyder, R. D.: Forecasting time series with complex seasonal patterns using exponential smoothing, J. Am. Stat. Assoc., 106, 1513–1527, 2011.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B, 1–38, 1977.

Donkor, E. A., Mazzuchi, T. A., Soyer, R., and Alan Roberson, J.: Urban water demand forecasting: review of methods and models, J. Water Res. Pl.-ASCE, 140, 146–159, 2012.

Ferraty, F. and Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics), Springer-Verlag New York, Inc., 2006.

Gaffney, S. and Smyth, P.: Trajectory clustering with mixtures of regression models, in: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 63–72, 1999.

Gaffney, S. and Smyth, P.: Curve Clustering with Random Effects Regression Mixtures, in: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AISTATS), 2003.

Gaffney, S. J.: Probabilistic curve-aligned clustering and prediction with regression mixture models, PhD thesis, University of California, Irvine, 2004.

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., and Meijers, E.: Smart Cities – Ranking of European medium-sized cities, Vienna University of Technology, 2007.

Gourieroux, C., Monfort, A., and Gallo, G.: Time series and dynamic models, vol. 3, Cambridge University Press, 1997.

Hernández, L., Baladrón, C., Aguiar, J. M., Carro, B., and Sánchez-Esguevillas, A.: Classification and clustering of electricity demand patterns in industrial parks, Energies, 5, 5215–5228, 2012.

Irwin, G., Monteith, W., and Beattie, W.: Statistical electricity demand modelling from consumer billing data, IEEE proceedings, Part C, Generation, transmission and distribution, 133, 328–335, 1986.

Jacques, J. and Preda, C.: Model-based clustering for multivariate functional data, Comput. Stat. Data An., 71, 92–106, 2014.

MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, 1, 281–297, 1967.

McKenna, S., Fusco, F., and Eck, B.: Water demand pattern classification from smart meter data, Procedia Engineering, 70, 1121–1130, 2014.

McLachlan, G. and Krishnan, T.: The EM algorithm and extensions, 382, John Wiley & Sons, 2007.

Nam, T. and Pardo, T. A.: Conceptualizing smart city with dimensions of technology, people, and institutions, in: Proceedings of the 12th annual international digital government research conference, ACM, 282–291, 2011.

Peng, J. and Müller, H.-G.: Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions, The Annals of Applied Statistics, 1056–1077, 2008.

Ramsay, J. O. and Silverman, B. W.: Functional Data Analysis, Springer Series in Statistics, Springer, 2nd Edn., 2005.

Samé, A., Chamroukhi, F., Govaert, G., and Aknin, P.: Model-based clustering and segmentation of time series with changes in regime, Advances in Data Analysis and Classification, 5, 301–321, 2011.

Samé, A., Noumir, Z., Cheifetz, N., Sandraz, A.-C., and Féliers, C.: Décomposition et classification de données fonctionnelles pour l'analyse de la consommation d'eau potable (in french), in: Extraction et Gestion des Connaissances (EGC) conference – Clustering and Co-clustering (CluCo) workshop, 2016.

Schwarz, G.: Estimating the dimension of a model, Ann. Stat., 6, 461–464, 1978.

Shumway, R. H. and Stoffer, D. S.: Time series analysis and its applications: with R examples, Springer Science & Business Media, 2010.

Sood, A., James, G. M., and Tellis, G. J.: Functional regression: A new model for predicting market penetration of new products, Marketing Science, 28, 36–51, 2009.

Wang, J.-L., Chiou, J.-M., and Mueller, H.-G.: Review of functional data analysis, arXiv preprint arXiv:1507.05135, 2015.